NASA Contractor Report 4523

KNOWLEDGE-BASED MACHINE INDEXING FROM NATURAL LANGUAGE TEXT: Knowledge Base Design, Development and Maintenance

Michael T. Genuardi
RMS Associates
Linthicum Heights, Maryland

Prepared for
NASA Scientific and Technical Information Program
under Contract NASw-4070

NASA
National Aeronautics and
Space Administration

Scientific and Technical
Information Program

1993

KNOWLEDGE-BASED MACHINE INDEXING FROM NATURAL LANGUAGE TEXT:
Knowledge Base Design, Development and Maintenance*

Michael T. Genuardi
NASA Center for AeroSpace Information, Linthicum Heights, MD 21090-2934

One strategy for machine-aided indexing (MAI) is to provide a concept-level analysis of the textual elements of documents or document abstracts. In such systems natural-language phrases are analyzed in order to identify and classify concepts related to a particular subject domain. The overall performance of these MAI systems is largely dependent on the quality and comprehensiveness of their knowledge bases. These knowledge bases function to (1) define the relations between a controlled indexing vocabulary and natural language expressions; (2) provide a simple mechanism for disambiguation and the determination of relevancy; and (3) allow the extension of a concept-hierarchical structure to all elements of the file. After a brief description of the NASA Machine-Aided Indexing system, concerns related to the development and maintenance of MAI knowledge bases are discussed. Particular emphasis is given to statistically-based text analysis tools designed to aid the knowledge base developer. One such tool, the Knowledge Base Building (KBB) program, presents the domain expert with a well-filtered list of synonyms and conceptually-related phrases for each thesaurus concept. Another tool, the Knowledge Base Maintenance (KBM) program, functions to identify areas of the knowledge base affected by changes in the conceptual domain, for example, the addition of a new thesaurus term. An alternate use of the KBM as an aid in thesaurus construction is also discussed.

1. Introduction

The primary goal of natural language processing (NLP) is to establish a machine system that can effectively determine the conceptual content of written text and manipulate those concepts in order to provide a response which mimics some human intellectual activity. Although this goal has not been achieved, many of the analysis methods developed in support of natural language research have been incorporated into operational systems that function as computer aids rather than as fully automatic techniques. In the area of machine-aided indexing (MAI), various strategies for statistical and syntactic analysis and knowledge base design have been used. The function of such machine-aided indexing systems is to provide a concept-level analysis of the textual elements of documents or document abstracts—the final output being a list of candidate index terms from an established classification scheme or thesaurus.

This paper will present an outline of the functional elements comprising knowledge-based MAI systems along with a description of a system currently in operation at the NASA Center for AeroSpace Information (CASI). Secondly, the development and maintenance of MAI knowledge bases are discussed with particular emphasis being given to statistically-based text analysis tools designed to aid the knowledge base developer. Finally, the application of these tools as an aid in thesaurus construction is described.

2. Machine-Aided Indexing through Text Analysis

Functional Elements

As inferred above, the types of MAI systems of concern here are those which function through the analysis of text. There are, of course, other strategies for providing assistance to the indexer. The MedIndEx Project[1] at the National Library of Medicine, for example, is currently developing an interactive expert system based on the "rules of indexing", where the indexer is guided through the indexing process in a somewhat heuristic manner.

The operations that comprise text-based MAI systems (Fig. 1) can be generalized as the following:

- delineation of text phrases
- identification/reduction of semantic units
- semantic analysis

The primary task of the first of these operations is to establish boundaries or parameters that will assure that two or more non-adjacent words, selected for subsequent semantic interpretation, actually represent a grammatically 'correct' association (such as between a modifying adjective and the noun it was intended to modify). This can be carried out by

* Revision of a paper presented at the Second International Congress on Terminology and Knowledge Engineering, Trier (Germany), 2-4 Oct., 1990

using techniques that vary in complexity from simple phrase-breaking procedures to formal syntactic parsing. Simple non-syntactic techniques have the advantage of being time-efficient; full parsing, on the other hand, may provide greater accuracy.

The second operation involves the extraction of multi-word strings (and single words) which may express concepts within a given subject domain. This operation typically requires a means for the successive concatenation of words within phrases and may optionally include a process for word-stemming (as in PHOTOELECTRICAL to PHOTOELECTRIC) or phrase normalization (as in SURFACE OF THE MOON to MOON SURFACE).

In the third operation, the final forms of the words and word-strings identified are then translated into appropriate indexing terms from the controlled vocabulary. One of the primary functions of the knowledge base is to serve as the equivalency table for this translation process.

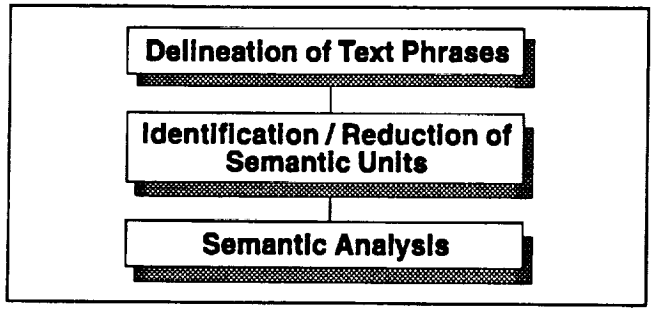

Figure 1. MAI Systems Operations

Knowledge Base Design and Function

The core of the MAI knowledge base is the thesaurus or classification scheme used for indexing. These controlled vocabularies represent the concepts of interest within a particular subject domain. The MAI knowledge base can be viewed as a conceptual network that (1) defines the relations between controlled thesaurus terms and natural language expressions, and (2) allows the extension of thesaurus hierarchical structure to all elements of the file.

Beside containing entries that map controlled terms to textual expressions, the knowledge base contains entries that represent decisions regarding the relevancy of particular concepts. For example, within an aeronautics domain, the concept AIRCRAFT is much too broad in meaning to be a relevant indexing term for most instances where the word *aircraft* appears in text. In this case, specific entries in the knowledge base would affect a search for a larger semantic unit (such as *aircraft stability, A-320 aircraft, aircraft construction materials*, etc.). Also, file entries are included that serve to disambiguate certain words; for example, whether the word *matrices* refers to mathematical matrices or material matrices.

Naturally, the form that any particular knowledge base takes on is dependent on how the other system operations are carried out. The procedures selected for initial phrase delineation and analysis define what kinds of information need to be represented in knowledge base entries and also how large an operational file will need to be (e.g., the use of word-stemming and phrase normalization can reduce the number of required entries). Likewise, the strategies used for disambiguation and relevancy analysis define the level of complexity required for knowledge representation and ultimately may dictate what kind of data structure is utilized.

NASA MAI System

The NASA Machine-Aided Indexing Project was initiated several years ago and had as its goal the development of two operational systems: one to carry out "subject switching"[2], (i.e., the translation of terms from one controlled vocabulary to the terms of another); and an MAI system based on the analysis of natural language text (Fig. 2). The designs for both of these were based on a "phrase structure rewrite" method, the historical development of which is described by Klingbiel (1985)[3]. Very simply, a phrase structure rewrite system, or "lexical dictionary", is a table format and access procedure that provides an efficient means for the translation of single and multi-word phrases to a controlled vocabulary.

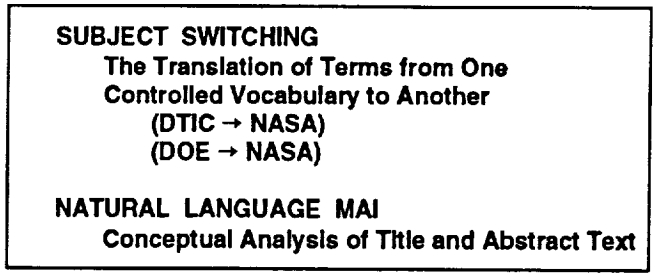

Figure 2. NASA MAI Project

Although the subject switching system was used in an operational setting, the application of the text system was limited due to its unacceptably-long response time when used in an interactive workstation environment. An additional problem was the slow development time and level of manual effort associated with knowledge base construction. In 1987, a re-design effort was initiated that focused on the evaluation of the phrase delineation process. The delineation process was based on the syntactic analysis of input text and thus required (1) that the syntactic class of each word be identified from a separate table, and (2) that the sequence of resulting syntactic classes be checked against a table of grammar rules.

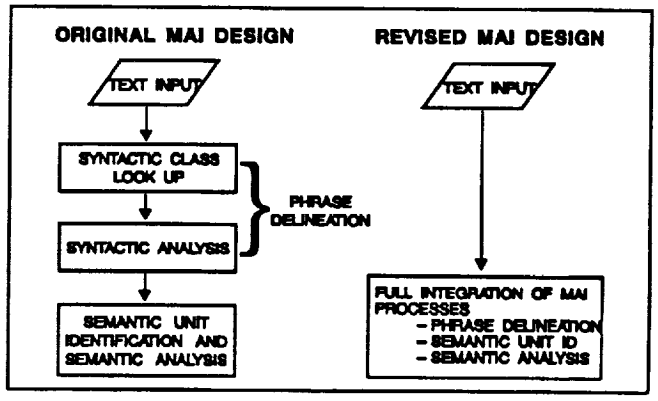

Figure 3. Design Comparison

In the new design the syntactic procedure was replaced by a simple preprocessing step and the incorporation of a proximity limit for words concatenated during semantic-unit identification. Preprocessing consists of breaking raw text input at end-punctuation (periods, colons, semicolons) and at the occurrence of certain 'stop-words.' The proximity limit is a constraint imposed on the word-concatenation process (that process functioning to identify semantic units for subsequent look-up in the knowledge base). It is an empirically established value above which the likelihood of grammatically 'incorrect' word associations becomes significant (thereby resulting in regular errors in the final MAI output).

This and other revisions of NASA MAI resulted in a system that was able to match and in some cases improve upon the output of the original system. In addition, response time was reduced by approximately 80 percent.

The NASA system represents a good example of MAI process integration. In the same way that the phrase delineation process was incorporated into the process for semantic-unit identification, the concatenation method used for this second process is integrated with the final MAI process—semantic-unit translation (i.e., the knowledge base look-up). The method for the identification of semantic-units is carried out through the execution of concatenation logic rules that dynamically incorporate information from knowledge base entries. Thus, in many instances, the existence of specific knowledge base entries directs the concatenation process.

3. Knowledge Base Development

Development Tools

As was mentioned earlier, the particular form and content of an MAI knowledge base is dependent on the design for carrying out the basic system operations. To a large extent, a system can compensate for design tradeoffs by incorporating the appropriate class of entries into the knowledge base. The NASA system, for example, does not include a mechanism for word-stemming; thus, all variant forms of words are included in knowledge base entries. One of the primary concerns of the system designer, then, becomes the tradeoff between the size of the knowledge base file and the response time associated with a particular level of system complexity.

Regardless of the specific design selected for an MAI system, the overall performance of these systems is largely dependent on the quality and comprehensiveness of their knowledge bases. Strict control and input from domain experts are critical during the development process. In fact, it may be the construction of the knowledge base itself which requires the greatest amount of development time and resources. The bulk of knowledge base content is repre-sented by the entries mapping natural language expressions to thesaurus concepts. The identification of those expressions may seem like an infinite task, especially if the thesaurus representing the subject domain is very large. KWIC (Key Word In Context) indexes of available text are of little use precisely because they are arrangements by 'key words' rather than the actual target concepts. Obviously, a review of abstract text on a case-by-case basis is grossly inefficient and is likewise untargeted with regard to domain concepts. In addition, both of these strategies lead to an unnecessarily large knowledge base due to the addition of expressions that are essentially 'unique,' i.e., text expres-sions that have a very low frequency of occurrence. A statistically-based text analysis tool was designed, again in support of the NASA project, that presents the domain expert with a well-filtered list of synonymous and conceptually-related phrases for each thesaurus concept. This tool was designed to satisfy three main requirements (Fig. 4):

- The output phrases would be targeted to one specific concept (i.e., the phrases considered during any particular session would be related to a single thesaurus term; thus, all expressions related to a particular domain concept could be analyzed together)

- The output phrases would be restricted to those that had a high frequency of occurrence within the existing NASA database (thus, 'unique' expressions would be screened out)

- The phrases would be in a file-compatible form (i.e., the phrases would be normalized to a form which could be extracted by the semantic-unit identification process)

Targeted, Meaningful Text	Candidate Synonyms for Each Thesaurus Term
File Compatible Phrases	Phrase Format as Recognized by Semantic ID Process
High Frequency of Occurrence	Unique Expressions Screened Out

Figure 4. Requirements for Text Analysis

The basic processing steps of the Knowledge Base Building Tool, or KBB can be described as follows:

1) The text used for input is comprised of the titles and abstracts from a large (150-2000) set of bibliographic records related to a single thesaurus concept—standard on-line search capabilities are used to identify an accurate set of records.

2) The text is copied into a file and preprocessed using a simple text-breaking method similar to that used in the MAI process.

3) A word-concatenation process is then used to identify all possible multi-word phrases within a maximum length (five). A proximity-limit for concatenation is imposed along with certain rules that provide syntactic filtering (which, for example, prevent prepositions and articles from beginning or ending a phrase).

4) A count of the frequency-of-occurrence is determined for each unique single-word and multi-word phrase. The words and phrases are then sorted in descending order by the frequency values. A lower-limit value is established, under

which all associated phrases are eliminated. However, there is a natural bias for single-words to have much higher frequencies than two-word phrases; which, in turn, will have higher frequencies than three-word phrases; etc. This can be dealt with in two ways. One is to simply produce five separate sorts, each one corresponding to a different phrase length. The other is to utilize a derived frequency value that effectively accounts for the bias. A process for determining such a value was recently described by Jones, Gassie, and Radhakrishnan (1990)[4]. The formula can be stated as $W F N^2$, where W is the sum of the frequencies of the words in the phrase, F equals the frequency of the phrase, and N equals the number of distinct words in a phrase. The N^2 was an empirically established relation found to be optimal for their particular application.

5) The final processing procedure serves to further refine the output. The phrases are checked against the existing knowledge base to eliminate any phrase that properly translates to a thesaurus concept other than the one that the KBB is currently analyzing; and to eliminate single-words and phrases that have a poor or low semantic value.

		Three-word phrase output			Single-word output
482	*	METAL MATRIX COMPOSITE(S)	74		FIBER-MATRIX
72		BEHAVIOR OF COMPOSITES	70	*	MMCS
72		STRENGTH OF COMPOSITE(S)	47		REINFORCEMENTS
62	*	REINFORCED METAL MATRIX	45		FIBER/MATRIX
61	*	ALUMINUM MATRIX COMPOSITE(S)	41	*	SIC/AL
55		PROPERTIES OF COMPOSITE(S)	29		STRENGTHENING
51		REINFORCED MATRIX COMPOSITE(S)	29		UNREINFORCED
49	*	REINFORCED METAL COMPOSITE(S)	27	*	BORON/ALUMINUM
48	*	REINFORCED ALUMINUM COMPOSITE(S)	25		MODULI
46		FIBER AND MATRIX	24	*	GRAPHITE/ALUMINUM
42	*	FIBER REINFORCED METAL(S)	21	*	AL-SIC
40		FIBER MATRIX COMPOSITE(S)	20		FP
38		BEHAVIOR OF MATRIX	19		STRENGTHENED
37		BEHAVIOR OF METAL	18		MICROGRAPHIS
35		FIBER REINFORCED MATRIX	17	*	AL-MATRIX
33	*	FIBER METAL COMPOSITE(S)	17	*	ARALL
33	*	FIBER REINFORCED ALUMINUM	16		ADDITIONS
33		PROPERTIES OF REINFORCED	16		EXTRUDED
32		PROPERTIES OF MATRIX	16		FRACTOGRAPHIC
31	*	FIBER METAL MATRIX	16	*	GR/AL
30		PROPERTIES OF METAL	16	*	GR/MG
29	*	ALUMINUM ALLOY MATRIX	16		PARTICULATE-REINFORCED
29		FIBER VOLUME FRACTION	16		SIC-REINFORCED
29		STRENGTH OF FIBER(S)	15	*	AL-SI
28	*	ALLOY MATRIX COMPOSITE(S)	14	*	AL/SIC
28	*	ALUMINUM ALLOY COMPOSITE(S)			
28		CHARACTERISTICS OF COMPOSITE(S)			
28		PROPERTIES OF ALUMINUM			
28		PROPERTIES OF FIBER(S)			
27	*	METAL MATRIX MATERIAL(S)			
26	*	SILICON CARBIDE ALUMINUM			
24		PROPERTIES OF ALLOY(S)			
24	*	SIC REINFORCED ALUMINUM			
23	*	ALUMINUM METAL MATRIX			
22		BEHAVIOR OF ALUMINUM			
22		HIGH TEMPERATURE COMPOSITES			
22	*	REINFORCED ALUMINUM ALLOY(S)			
22		SILICON CARBIDE WHISKER(S)			
22		TRANSMISSION ELECTRON MICROSCOPY			
21	*	FIBER ALUMINUM COMPOSITE(S)			
21		THERMAL EXPANSION COEFFICIENT(S)			
20	*	CARBIDE REINFORCED ALUMINUM			
20		FATIGUE CRACK GROWTH			
20		RULE OF MIXTURES			
20		SCANNING ELECTRON MICROSCOPY			
20		STRENGTH AND FRACTURE			

Table 1. Un-edited KBB output for METAL MATRIX COMPOSITES

Sample output form the Knowledge Base Building Tool (KBB) is shown in Table 1. The input consisted of titles and abstracts from records associated with the thesaurus concept METAL MATRIX COMPOSITES. The first column in this table list the unedited three-word phrase output. Those phrases selected by a subject analyst for inclusion in the knowledge base are indicated with asterisks (*). The second column lists the output that the KBB program identified as being single words. Several acronyms and material abbreviations have been recognized and flagged by a subject analyst.

Maintenance Tools

Statistical text analysis procedures like the KBB are interesting in that, an analysis of their output suggests many possible alternate applications. One problem associated with the maintenance of MAI knowledge bases, arises from the fact that the conceptual domain, as represented by the controlled vocabulary, is not static. New terms are regularly added and integrated into existing conceptual hierarchies. The areas of the knowledge base requiring modification in response to such changes cannot be easily inferred — particularly if the file happens to be very large (the NASA file currently has over 113,000 entries).

A modification of the KBB program, the KBM (Knowledge Base Maintenance) routine, provides a tool for the identification of affected entries. The final procedure in the KBB process was altered to allow phrases already translating to a thesaurus term to be included in the final output and flagged for easy recognition. Text expressions which may been mapped to an existing thesaurus term in lieu of the newly established term will be evident in the program output.

4. Other Applications and Future Directions

The KBM program has more recently been used as an aid to thesaurus construction. The phrase output provides some guidance in identifying trends in the lexicon of the particular subject area; the existing thesaurus terms associated with the identified phrases suggest probable hierarchy locations and related terms for new thesaurus entries. The program is particularly useful when investigating an emerging technology or discipline. Sample phrase output for the general area of ROBOTICS is shown in the first column of Table 2. The second column presents some output phrases that have been selected and conceptually organized by a lexicographer.

Since the particular text input to the KBM is identified through a traditional on-line database search, the conceptual scope of the text to be processed can be easily modified. A separate corpora of input text associated with the more specific area of ROBOT VISION was processed in conjunction with the analysis of general ROBOTICS. Sample output is shown in Table 3. Synonyms have been flagged with "=" by a lexicographer and other phrases relevant to the lexicon of this particular discipline have been flagged with (*).

The limitation of the type of text-based MAI system described in this paper is that the semantic unit is restricted to

the level of the phrase. These systems are best suited to document indexing applications associated with technical and scientific domains, where the likelihood of a phrase containing specific indexable content is significant. In such environments they provide both quality and production benefits. Some interesting possibilities exist for the application of these MAI systems to a full-text environment. The same basic design could be modified to capture occurrence frequencies of suggested thesaurus terms. A term weighting scheme could be developed that incorporated these statistical values and special weight values assigned to terms originating from key structural elements of the document, such as title, section headings, abstracts, etc. Such a system may very well provide a quality of output that would allow its use as a totally automatic indexing system.

172	ROBOT MANIPULATOR(S)	ROBOT DYNAMICS
141	ROBOT ARM(S)	ROBOT MOTION
110	ROBOTIC MANIPULATOR(S)	MANIPULATOR DYNAMICS
99	ROBOTIC SYSTEM(S)	..TRAJECTORY PLANNING
91	ROBOT CONTROL	..PATH PLANNING
90	MOBILE ROBOT(S)	..MOTION PLANNING
84	ROBOT SYSTEM(S)	..DYNAMIC CONTROL
67	CONTROL SCHEME(S)	..POSITION CONTROL
65	END EFFECTOR(S)	..OBSTACLE AVOIDANCE
63	CONTROL ALGORITHM(S)	..INVERSE KINEMATICS
61	MANIPULATOR SYSTEM(S)	
55	CONTROL LAW(S)	ROBOT JOINTS
53	COMPUTER SIMULATION(S)	
52	INVERSE KINEMATIC(S)	ROBOT ARMS
49	CONTROL ROBOT(S)	MANIPULATOR ARMS
49	VISION SYSTEM(S)	ROBOTIC MANIPULATORS
47	FORCE CONTROL	
46	FLEXIBLE ARM(S)	ROBOT HANDS
44	CONTROL PROBLEM(S)	END EFFECTORS
44	CONTROL STRATEGY(IES)	
44	SPACE ROBOTIC(S)	ROBOT SENSORS
42	AUTOMATION ROBOTICS	..TACTILE SENSORS
41	AUTONOMOUS ROBOT(S)	..TACTILE SENSING
40	MANIPULATOR ARM(S)	..TORQUE SENSORS
39	MANIPULATOR CONTROL	..FORCE CONTROL
37	DYNAMIC ROBOT	
37	ROBOT DYNAMICS	ROBOT VISION
35	CONTROL METHOD(S)	MACHINE VISION
35	DYNAMIC MODEL	COMPUTER VISION
35	MOTION CONTROL	VISION SYSTEMS
35	TELEROBOTIC SYSTEM(S)	
34	FLEXIBLE MANIPULATOR(S)	ROBOT CONTROL
33	AUTONOMOUS VEHICLE(S)	
32	ADAPTIVE CONTROLLER(S)	ROBOTS
32	FEEDBACK CONTROL	..INDUSTRIAL ROBOTS
32	INDUSTRIAL ROBOT(S)	..SPACE MANIPULATORS
31	PATH PLANNING	..SPACE TELEROBOTICS
30	CONTROL MANIPULATORS	
30	DYNAMIC MANIPULATORS	
30	ROBOTIC APPLICATIONS	
30	SPACE TELEROBOTIC(S)	
29	POSITION CONTROL	
28	CONTROL CONTROL	
27	INTELLIGENT ROBOT(S)	
27	ROBOT MOTION	
27	SPACE ROBOT(S)	
26	DYNAMIC CONTROL	
26	DYNAMIC SYSTEM(S)	
25	FLEXIBLE ROBOT	
25	INTELLIGENT SYSTEM(S)	
25	OBSTACLE AVOIDANCE	
24	MODEL CONTROL	
23	ASSEMBLY TASKS	
23	DYNAMIC EQUATIONS	
23	MOTION PLANNING	
23	ROBOT HANDS	
23	TRAJECTORY PLANNING	
	Un-edited output from general text on ROBOTICS (two-word phrases)	Select output phrases conceptually organized

Table 2. KBM applied to thesaurus construction—output from text related to a broad subject field (ROBOTICS)

205	=	COMPUTER VISION
191		VISION SYSTEM(S)
60	=	MACHINE VISION
54	*	OBJECT RECOGNITION
46	=	ROBOT VISION
38	*	IMAGE ANAYLSIS
37		ROBOTIC SYSTEM(S)
35	*	PATTERN RECOGNITION
33		ROBOT SYSTEM(S)
33	*	VISION ALGORITHM(S)
30	*	FEATURE EXTRACTION
30	*	SCENE ANALYSIS
28	*	EDGE DETECTION
27	*	STEREO VISION
27		VISUAL SYSTEM(S)
23		RANGE DATA
22	*	IMAGE SEGMENTATION
22		MOBILE ROBOT
22	=	ROBOTIC VISION
22		THREE-DIMENSIONAL OBJECTS
20		IMAGE DATA
20		RECOGNITION SYSTEM
19	*	LOW-LEVEL VISION
18	*	IMAGE FEATURES
18		SPACE APPLICATIONS
18	*	SURFACE ORIENTATION
18		VISION APPLICATIONS

Table 3. KBM applied to thesaurus construction—un-edited output from text related to a narrow subject field (ROBOT VISION)

References

(1) HUMPHREY, S.M. and MILLER, N.E. Knowledge-based indexing of the medical literature. The indexing aid project. JASIS 38 (1987), p. 184-196.

(2) SILVESTER, J.P., NEWTON, R., KLINGBIEL, P.H. An operational system for subject switching between controlled vocabularies: A computational linguistics approach. NASA-CR-3838, (1984).

(3) KLINGBIEL, P.H. Phrase structure rewrite systems in information retrieval. Information Processing and Management 21 (1985), no. 2, p. 113-126.

(4) JONES, L.P., GASSIE, E.W., RADHAKRISHNAN, S. INDEX: The statistical basis for an automatic conceptual phrase-indexing system. JASIS 41 (1990), no. 2, p. 87-97.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (leave blank)	2. REPORT DATE March 1993	3. REPORT TYPE AND DATES COVERED Contractor Report
4. TITLE AND SUBTITLE Knowledge-Based Machine Indexing from Natural Language Text: Knowledge Base Design, Development and Maintenance		5. FUNDING NUMBERS
6. AUTHOR(S) Michael T. Genuardi		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Center for AeroSpace Information Linthicum Heights, MD 21090-0234		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546		10. SPONSORING/MONITORING AGENCY REPORT NUMBER NASA-CR-4523

11. SUPPLEMENTARY NOTES
Presented at the Second International Congress on Terminology and Knowledge Engineering, Trier, Germany, 2-4 October 1990

12a DISTRIBUTION/AVAILABILITY STATEMENT Unclassified - Unlimited Subject Category - 82	12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

One strategy for machine-aided indexing (MAI) is to provide a concept-level analysis of the textual elements of documents or document abstracts. In such systems, natural-language phrases are analyzed in order to identify and classify concepts related to a particular subject domain. The overall performance of these MAI systems is largely dependent on the quality and comprehensiveness of their knowledge bases. These knowledge bases function to (1) define the relations between a controlled indexing vocabulary and natural language expressions; (2) provide a simple mechanism for disambiguation and the determination of relevancy; and (3) allow the extension of concept-hierarchical structural to all elements of the file. After a brief description of the NASA Machine-Aided Indexing system, concerns related to the development and maintenance of MAI knowledge bases are discussed. Particular emphasis is given to statistically-based text analysis tools designed to aid the knowledge base developer. One such tool, the Knowledge Base Building (KBB) program, presents the domain expert with a well-filtered list of synonyms and conceptually-related phrases for each thesaurus concept. Another tool, the Knowledge Base Maintenance (KBM) program, functions to identify areas of the knowledge base affected by changes in the conceptual domain (for example, the addition of a new thesaurus term). An alternate use of the KBM as an aid in thesaurus construction is also discussed.

14. SUBJECT TERMS computer techniques, information retrieval, knowledge bases (artificial intelligence), thesauri, terminology, natural language processing			15. NUMBER OF PAGES 8 16. PRICE CODE A02
17. SECURITY CLASSIFICATION OF REPORT Unclass	18. SECURITY CLASSIFICATION OF THIS PAGE Unclass	19. SECURITY CLASSIFICATION OF ABSTRACT Unclass	20. LIMITATION OF ABSTRACT Unlimited

Available from NASA Center for AeroSpace Information
800 Elkridge Landing Road
Linthicum Heights, MD 21090-2934
(301) 621-0390